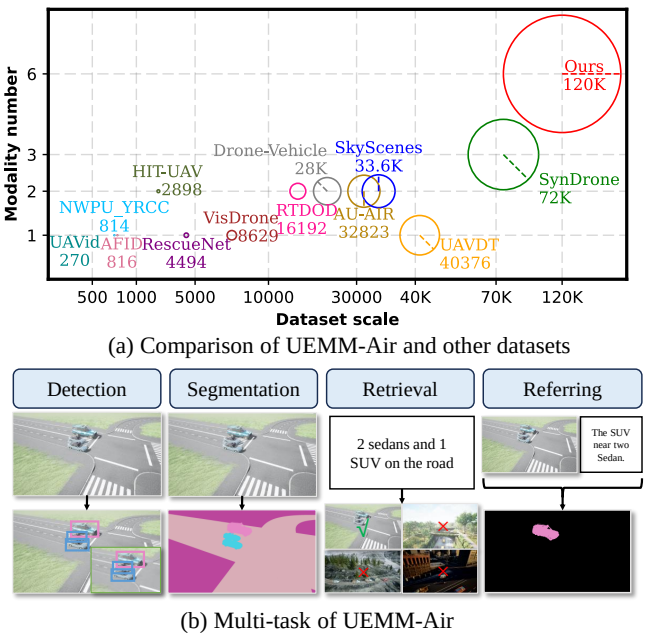
UEMM-Air: Make Unmanned Aerial Vehicles Perform More Multi-modal Tasks

Liang Yao, *Graduate Student Member, IEEE*, Fan Liu, *Member, IEEE*, Shengxiang Xu, Chuanyi Zhang, *Member, IEEE*, Xing Ma, Jianyu Jiang, Zequan Wang, Shimin Di, *Member, IEEE*, and Jun Zhou, *Senior Member, IEEE*

***Abstract*—The development of multi-modal learning for Unmanned Aerial Vehicles (UAVs) typically relies on a large amount of pixel-aligned multi-modal image data. However, existing datasets face challenges such as limited modalities, high construction costs, and imprecise annotations. To this end, we propose a synthetic multi-modal UAV-based multi-task dataset, UEMM-Air. Specifically, we simulate various UAV flight scenarios and object types using the Unreal Engine (UE). Then we design the UAV's flight logic to automatically collect data from different scenarios, perspectives, and altitudes. Furthermore, we propose a novel heuristic automatic annotation algorithm to generate accurate object detection labels. Finally, we utilize labels to generate text descriptions of images to make our UEMM-Air support more cross-modality tasks. In total, our UEMM-Air consists of *120k* pairs of images with *6* modalities and precise annotations. Moreover, we conduct numerous experiments and establish new benchmark results on our dataset. We also found that models pre-trained on UEMM-Air exhibit better performance on downstream tasks compared to other similar datasets. The dataset is publicly available (https://github.com/1e12Leon/UEMM-Air) to support the research of multi-modal tasks on UAVs.**

***Index Terms*—Unmanned Aerial Vehicles, Large Scale Dataset, Multi-modal, Multi-task**



(a) Comparison of UEMM-Air and other datasets

(b) Multi-task of UEMM-Air

Fig. 1. (a) Comparison of UEMM-Air and other UAV environmental perception datasets. (b) Various UEMM-Air targeted tasks. Our stands out as the largest in terms of data scale, featuring the most paired modality types and the greatest variety of tasks among existing datasets.

I. Introduction

With the advancement of Unmanned Aerial Vehicles (UAV) technology [1], [2], [3], [4], [5] and deep learning [6], [7], vision perception tasks of UAV have shown great potential in many fields such as urban monitoring, military reconnaissance and rescue [5], [8], [9], [10]. Unlike general vision task [11], UAV tasks exhibit characteristics such as complex backgrounds, varying scales, and small objects. Therefore, models trained on general datasets [12], [13] can hardly be directly applied to UAV platforms. To this end, many scholars have constructed vision datasets from the perspective of UAV. For example, VisDrone [14] and UAVDT [15] consider various scenes, weather conditions, and environments, providing good benchmarks for UAV-based Object Detection (UAV-OD) [16], [17], [18], [19] tasks. SkyScenes [20] encompassing different layouts, weather, and daytime conditions with corresponding dense annotations and viewpoint metadata for UAV-based Segmentation (UAV-Seg) [21], [22]. The aforementioned datasets make significant contributions to traditional UAV vision tasks.

However, with the development of multi-modal learning [23], [24], the above datasets are facing challenges such as being single-modality and having insufficient data. Therefore, the research communities are gradually moving from single-modal to multi-modal tasks. For instance, DroneVehicle [25] utilizes two image modalities: infrared and visible, with the infrared modality enhancing detection accuracy in nighttime scenes. However, due to the request for manual labeling and aligning two modalities images, the dataset annotation cost is relatively high. Although AU-AIR dataset [26] takes into account the potential value of UAV parameters, it covers relatively only a few scenes and its labeled samples have some imprecise annotations. Drawbacks in above datasets are adverse to model training. Furthermore, all the existing datasets are unable to support alignment of vision and lan-

Corresponding author: Fan Liu (fanliu@hhu.edu.cn).

Fan Liu, Liang Yao, Shengxiang Xu, Xing Ma, Jianyu Jiang, and Zequan Wang are with the College of Computer Science and Software Engineering, Hohai University, Nanjing, 210098, China.

Chuanyi Zhang is with the College of Artificial Intelligence and Automation, Hohai University, Nanjing, 210098, China.

Shimin Di is with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon Hong Kong.

Jun Zhou is with the School of Information and Communication Technology, Griffith University, Nathan, Queensland 4111, Australia.

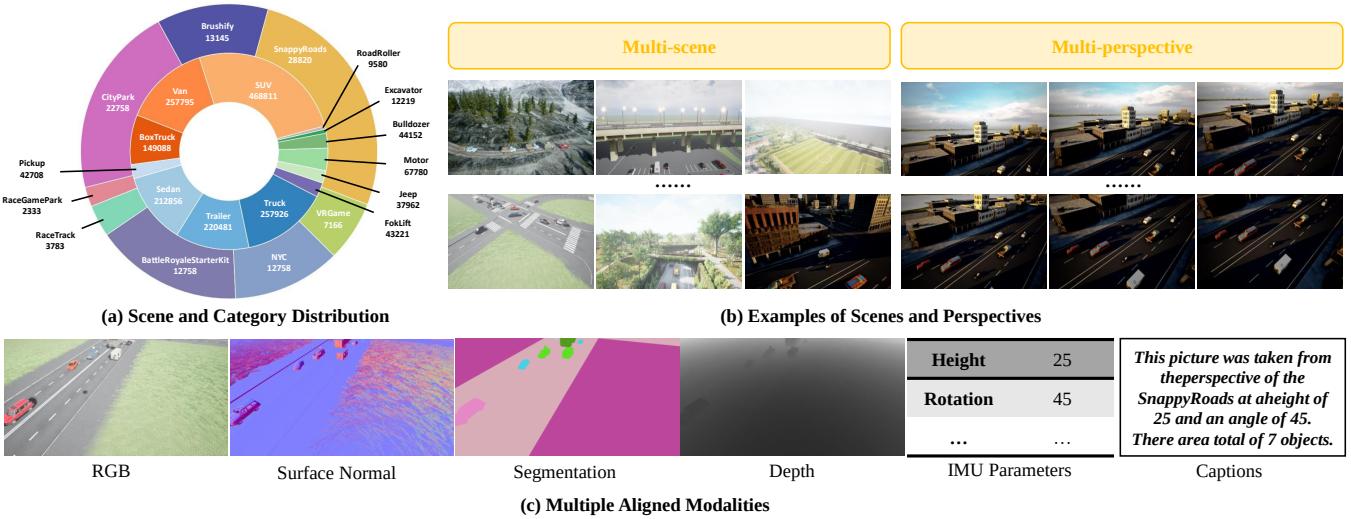


Fig. 2. UEMM-Air is a multi-scene, multi-modal, and multi-perspective UAV-based perception dataset. (a) Scence (outer) and object category (inner) distribution. (b) UEMM-Air features multiple scenes and various perspectives of the same view. (c) UEMM-Air encompasses 6 modalities: RGB, surface normals, segmentation, depth, IMU parameters, and textual descriptions.

guage to perform zero-shot inference like the CLIP [27] model in the field of UAVs. Similar to satellite remote sensing, UAVs also need numerous vision-language applications such as open-vocabulary object detection [28], referring image segmentation [29], and multi-modal large language models (MLLMs) [30], etc.

Motivated by the above observations, we construct a new large-scale synthetic UAV environmental perception dataset, UEMM-Air, to facilitate further research on the UAV field. As represented in Fig. 1 and Fig. 2, we assign UEMM-Air three significant characteristics. **(1) Multi-modal:** our dataset contains six modalities, including visible, depth, segmentation, surface normals, UAV IMU parameters, and captions. **(2) Multi-task:** our dataset is capable of supporting tasks such as object detection, instance segmentation, image-text contrastive learning, and referring image segmentation, etc. **(3) Multi-semantic:** our dataset covers a variety of scenarios and perspective, and has fine-grained category information.

To achieve these objectives, we first utilize the Unreal Engine (UE) [31] and AirSim [32] framework to build various simulated scenarios for UAV flights. Subsequently, we implement automatic UAV flight control and collect data at different altitudes, scenes, and modalities. Furthermore, we design an annotation algorithm to automatically generate object detection labels. Finally, we generate text descriptions for different cross-modal tasks according to existing detection and segmentation annotations. We also implement representative and impressive methods to systematically investigate the potential and challenges brought by UEMM-Air. The experimental results demonstrate the research significance of our dataset across various tasks.

The main contributions of this article are as follows:

- We propose a new synthetic UAV-based environmental perception dataset, UEMM-Air. To the best of our knowledge, **it is the largest dataset in terms of existing data scale, featuring the most paired modality types and the highest variety of task types.**
- We introduce a new heuristic algorithm for automatic data annotation. Compared with labeling strategies in other synthetic datasets, ours can provide more accurate annotations by introducing segmentation and depth modalities to enhance the identification of objects, especially in addressing visually overlapped objects.
- We conduct experiments on multiple types of tasks in the field of UAV, providing new benchmark results. Our UEMM-Air expands the field of UAV environmental perception from purely visual to multi-modal tasks.

The remainder of this paper is structured as follows. In Section II, we review related literature on UAV environmental perception and existing datasets. Section III introduces our UEMM-Air. Specifically, Section III-A introduces the automatic data collection methods we used. Then, in Section III-B, we discuss the configurations and advantages of each modality. We also outline the strategies for automatic annotation generation and the methods for cross-modal generation, respectively in Sections III-C and Section III-D. Next, Section IV validates the accuracy of our annotations and demonstrates the benchmarks of our UEMM-Air across multiple tasks, including object detection, instance segmentation, referring segmentation, and cross-modal retrieval. Finally, Section V presents our discussion of the advantages and limitations of our UEMM-Air and concludes this paper.

II. Related Work

A. UAV-based Environmental Perception

UAV-based environmental perception primarily involves two tasks: object detection and semantic segmentation.

1) Object Detection: Due to the typically top-down perspective of UAVs and the relatively small scale of the objects, general object detection methods [43], [44], [45], [46], [47] tend to be not suitable for UAV-based Object Detection (UAV-OD) tasks. The mainstream methods for UAV-based object

TABLE I
COMPARISON OF DIFFERENT UAV-BASED DATASETS. 'DET': OBJECT DETECTION, 'SEG': SEMANTIC SEGMENTATION, 'REF': REFERRING IMAGE SEGMENTATION, 'CL': IMAGE-TEXT CONTRASTIVE LEARNING. 'MM': MULTI-MODAL. 'ANGLE': UAV'S PAN&TILT VIEW ANGLE. '-': NOT APPLICABLE OR NOT EXPLICIT IN THEIR PAPERS.

Tasks	Dataset	Year	MM	# modalities	# images	# classes	Size [px]	Angle
Det	Stanford-Drone [33]	2016	✗	-	-	7	1450×1080	90
	UAVDT [15]	2018	✗	-	40376	3	1080×540	variable
	VisDrone [14]	2018	✗	-	8629	10	1920×1080	variable
	AU-AIR [26]	2020	✓	2	32823	8	1920×1080	45 to 90
	Drone-Vehicle [34]	2022	✓	2	28k	5	640×512	90
	HIT-UAV [35]	2023	✓	2	2898	4	640×512	30 to 90
	RTDOD [36]	2023	✓	2	16192	10	1280×720	variable
	State-Air [37]	2024	✓	2	2864	4	1280×720	variable
Seg	UAVid [38]	2020	✗	-	270	8	3840×2160	45
	NWPU_YRCC [39]	2020	✗	-	814	3	1600×640	variable
	RescueNet [40]	2022	✗	-	4494	10	3000×4000	variable
	AFID [41]	2023	✗	-	816	8	2560×1440	variable
	SkyScenes [20]	2024	✓	2	33.6k	28	2160×1440	variable
Dets, Seg	SynDrone [42]	2023	✓	3	60k	9	1920×1080	30,60,90
Dets, Seg, Ref, CL	**UEMM-Air (Ours)**	2025	✓	**6**	**120**k	**13**	1920×1080	variable

detection primarily employ coarse-to-fine strategies [48], [49], [50]. Initially, the detection process focuses on identifying larger objects, while concurrently pinpointing dense subregions containing small objects. These subregions are subsequently utilized as inputs for the model to refine detection results. For example, a CZDetector [51] employed a density-based cropping algorithm to identify regions with crowded objects and then increased the size of those regions to enhance the training dataset. Alternatively, [52] utilizes a Gaussian mixture model to supervise the detector in generating object clusters composed of focusing regions. To address limited computing resources, methods were proposed to balance accuracy and efficiency. Typical examples include CEASC [53] and SIFDAL [37]. The former adopted a plug-and-play detection method with enhanced sparse convolution and an adaptive mask scheme. The latter disentangled scale-invariant features to boost detection accuracy and mildly reduce test inference costs. Additionally, to adapt to the low computational power conditions of UAVs, some researchers [54], [55] employed compress techniques such as pruning and distillation.

2) Semantic Segmentation: Similar to object detection, there are many semantic segmentation methods specifically designed for UAV scenarios. High-resolution representation learning plays a crucial role in UAV semantic segmentation due to the ultra-high resolution and varying object scale of UAV remote sensing imagery [56], [57], [58]. For example, HRNet [59] was designed by a high-resolution network that repeatedly exchanges semantic information across adjacent multi-resolution sub-networks. At the same time, many lightweight methods were proposed to meet the real-time application requirements of UAV platforms [60], [61], [62]. Literature [63] also proposed Semantics Guided Bottleneck Network (SGBNet) based on BiSeNet and the Channel Pooled Attention (CPA) mechanism to balance segmentation accuracy, model size, and inference speed on the Land Cover Dataset.

B. UAV-based Environmental Perception Datasets

Many UAV-based environmental perception datasets provide multi-class images and videos captured by UAVs. These datasets are significant for promoting the research and development of various computer vision tasks, including object detection, object tracking, and scene understanding. We summarized several commonly used UAV-based datasets in Table I.

Stanford-Drone [33] is a large-scale dataset containing overhead images and videos of multi-class objects moving and interacting at Stanford University. This dataset can be used for learning and evaluating multi-object tracking, activity understanding, and trajectory prediction.

UAVDT [15] has 80,000 representative frames which are annotated with bounding boxes and 14 kinds of attributes in various complex scenarios. It focuses on 3 specific computer vision tasks: object detection, single-object tracking, and multiple-object detection.

VisDrone [14] is a large-scale benchmark dataset in object detection and tracking with various environment conditions and camera viewpoints. It contains 10 categories objects of frequent interest in drone applications and more than 2.5 million annotation bounding boxes.

AU-AIR [26] includes extracted frames meta-data, bounding box annotations for traffic-related object categories, and multi-modal flight sensor data. The dataset is captured at low altitudes at the intersection.

Drone-Vehicle [34] offers a drone-based RGB-Infrared cross-modality vehicle detection dataset and corresponding precise annotations. This dataset covers multiple scenarios and objects from day to night with three different angles and heights.

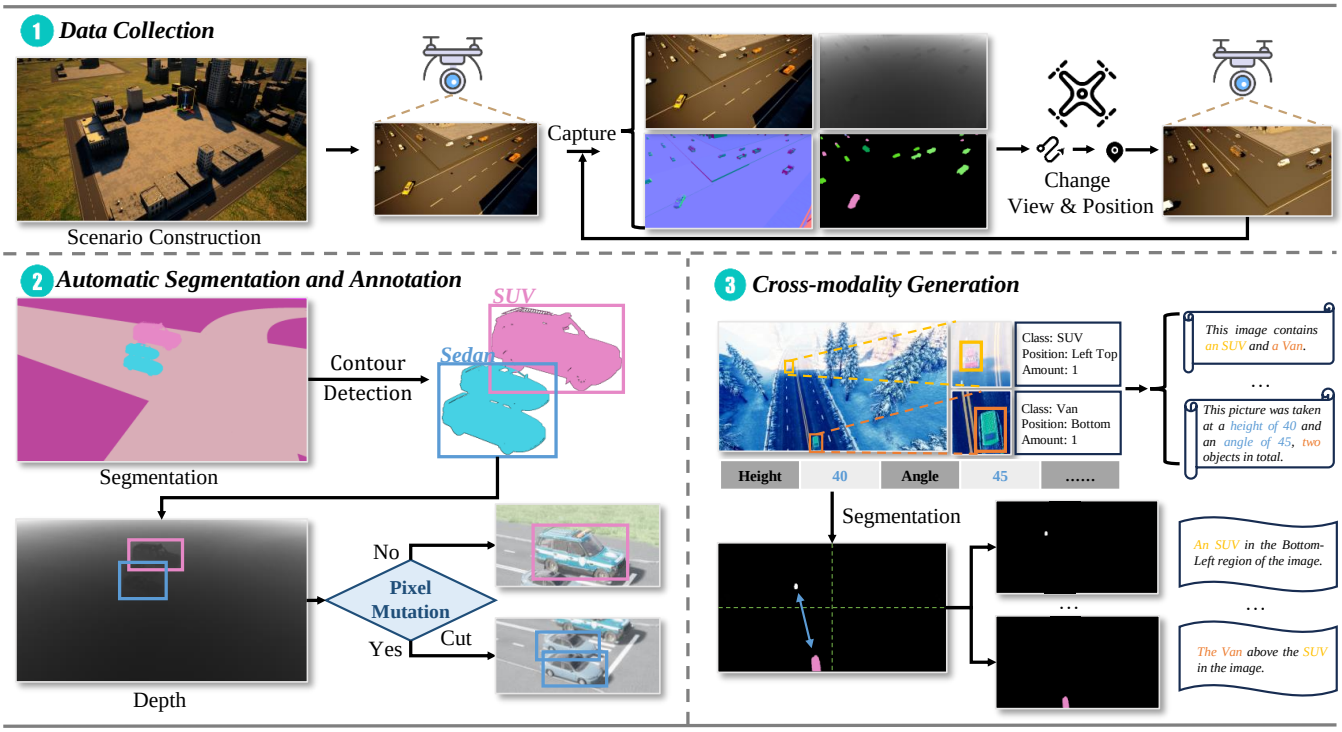


Fig. 3. Pipeline of our data construction. **Step 1:** We design the automatic flight logic of the UAV to collect images from different altitudes, perspectives, and modalities. **Step 2:** We perform contour detection on the segmentation image to obtain object bounding boxes. To alleviate visually overlapped situations, we introduce the depth information, where a significant change in depth typically indicates multiple objects. **Step 3:** We extract the objects' categories, quantities, and spatial relationships to generate captions for image-text contrastive learning and referring image segmentation.

HIT-UAV [35] is the first publicly available high-altitude UAV-based infrared thermal dataset for object detection applications on UAVs. The dataset provides two types of bounding box annotations (oriented and standard) to tackle the challenge of overlapping object instances in aerial images.

RTDOD [36] is the first RGB-Thermal domain-incremental object detection dataset designed specifically for UAVs. The dataset covers a wide range of weather conditions and day-to-night transitions, highlighting the complexity of real-world scenarios.

State-Air [37] is an aerial dataset with multi-modal sensor data collected in real-world outdoor environments. The dataset concludes 2246 images of sunny days and 616 instances of snowy ones with 4 categories: person, car, bus, and van.

UAVid [38] is a semantic segmentation dataset designed for UAV semantic segmentation in complex urban scenes, featuring on both static and moving object recognition. It provides 300 high-resolution oblique-view UAV images, labeled with 8 classes, and gives a diverse representation of objects with rich scene context.

NWPU_YRCC [39] is the first public UAV image dataset containing 814 accurately annotated images for river ice segmentation. It covers typical images of river ice in different periods, with diverse appearances and captured from different flight heights and views.

RescueNet [40] propose a high-resolution post-disaster dataset for natural disaster damage assessment. It includes detailed classification and semantic segmentation annotations, enabling applications in building damage classification, road segmentation, and future disaster assessment.

AFID [41] is a publicly available dataset of aerial and fluvial images, featuring detailed semantic annotation from different camera perspectives. It focuses on utilizing semantic segmentation models to fulfill unmanned hydrologic data collection, environmental inspection, and disaster warning tasks.

SkyScenes [20] is a synthetic dataset of densely annotated aerial images captured from UAV perspectives, containing 33.6K images from different altitudes and pitch settings. It provides pixel-level semantic, instance, and depth annotations, and enables reproducing the same viewpoint under different conditions.

SynDrone [42] proposes a multi-modal synthetic benchmark dataset containing both images and 3D data taken at multiple flying heights. It includes 28 classes of pixel-level labeling and object-level annotations for semantic segmentation and object detection.

III. UEMM-Air

A. Scene Construction and Flight Control Logic

Previous UAV-based datasets are limited in scene diversity, which tends to affect model generalization. Therefore, we aim to construct a dataset with richer scenes to improve the performance of the models. To be specific, we utilize Unreal Engine [31] with CityBLD [64] plugin. It can create cities of almost any size and style in a very short time to simulate scenarios in the real world. We build several scenes including

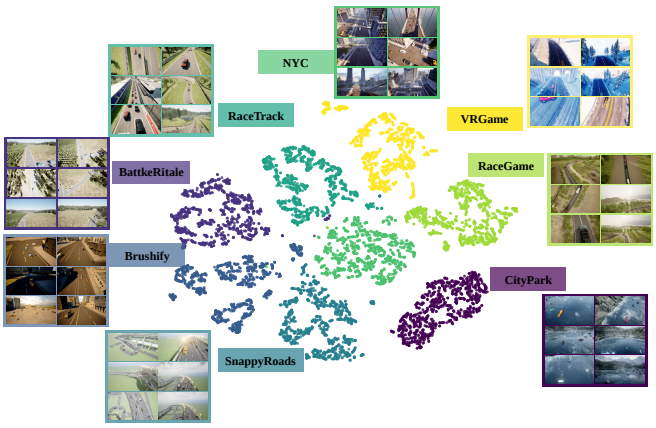


Fig. 4. We randomly sampled images from various scenes and visualized the features extracted by the CLIP image encoder through T-SNE. The significant differences in features across different scenes indicate that our dataset is beneficial for enhancing the model's domain generalization performance.

cities, parks, highways, etc. We also collect a total of 13 categories and more than hundreds of vehicle models.

We leverage Unreal Engine's movement animation to simulate dynamic scenes in reality. Employing the traffic features of Unreal Engine, we can flexibly design and construct various complex road layouts, including city streets, highways, and country roads. These layouts can precisely simulate real-world terrain and traffic conditions, providing realistic infrastructure for game scenes. Additionally, we can generate a wide variety of vehicles in the virtual environment. These vehicles can automatically navigate the generated roads based on predefined traffic rules and behaviors. By setting paths and control parameters, the vehicles can simulate real traffic flow, obey traffic signals, avoid pedestrians, and respond to traffic congestion, thereby creating highly realistic dynamic traffic scenarios.

To collect data, we control the UAV to fly and take pictures in Unreal Engine, as represented in Fig. 3 Step 1. Specifically, we build an Unmanned Aerial Vehicle simulator using AirSim and Pygame. When flying to a satisfactory shooting point, we control the UAV to fly within a height range of 5 meters to 50 meters above the horizontal surface, taking a set of photos every 5 meters up. The camera rotates from 0 degrees to 90 degrees by 5 degrees for each step. To obtain aligned pictures of different modalities, our simulator temporarily stops running when taking photos.

Ultimately, we built 8 different maps and collected a total of $120k$ pairs of data. Each map has its own unique style, incorporating various elements such as different lighting, scenes, and weather conditions. As shown in Fig. 4, we randomly select 1,000 images from each map and utilize CLIP to extract image features for T-SNE visualization analysis. The results indicate considerable feature differences among different maps, providing rich domain representations for model training.

B. Acquisition Setup

We adopt our camera sensor setup for the AirSim simulator to ensure diversity in data. The acquisition pipeline equips the UAV with several co-registered sensors. With the help of these sensors, we collect 5 modalities: RGB, infrared, segmentation, surface normal, and IMU parameters.

RGB Camera: It offers a resolution of 1920×1080 pixels. The vertical field of view (FoV) increases dynamically from $0°$ to $90°$, indicating that the viewing angle changes from a horizontal to a top-down view. All RGB images are stored in PNG format.

RGB images contain rich color, and spatial information, facilitating better image understanding and object recognition. The visual image is the most common modality in computer vision tasks. However, in complex environments such as nighttime, visual images alone may not perform well due to the poor visibility and the resulting inability to effectively detect objects.

Depth Camera: The depth camera has the same FoV, resolution and storage format as the RGB camera. It interpolates each pixel value from 0 to 255 according to the depth of the distance from the camera plane. The white pixels show a depth of more than 100 meters, while the black pixels indicate a depth of 0 meters.

Depth images leverage pixels to represent the distance from the object to the camera, reflecting the spatial shape and structure of the photographed scene. Therefore, we utilize it to address the issue of inaccurate annotations caused by overlapped visual information of the objects during the annotation generation process. It can also be leveraged to deduce an object's height, convexity, and relative position, which aids in multi-modal object objection.

Segmentation Camera: The segmentation camera maintains the same FoV, resolution, and data format as the preceding cameras. It generates distinct colors for pixels belonging to different categories of objects to ensure accurate segmentation of the scene.

The segmentation image divides the image into multiple regions with similar attributes, providing pixel-level information where each pixel is assigned to a precise category label. Because of the detailed segmentation information, this modality can assist in the automatic generation of detection annotations. Additionally, since segmentation images inherently contain positional information, combining them with other modalities for detection often leads to improved accuracy.

Surface Normal Camera: The Surface Normal Camera maps the X, Y, and Z components of the surface normal to an RGB range from 0 to 255. Due to the gradual changes in normal direction, it is difficult to distinguish. Therefore, the contrast of the normal camera images is set as 1.5 to more distinctly delineate changes in the direction of the normals. This camera saves its pictures in PNG format and has the same field of view and resolution as the Scene Camera.

Surface normal images primarily capture the geometric features and surface details of the target object. When fused with RGB images or other modalities, they can compensate for deficiencies in texture features. For example, in fine-grained object detection, the texture features introduced by the surface normal modality can help the model learn deeper fine-grained information. Additionally, it reveals intricate surface details

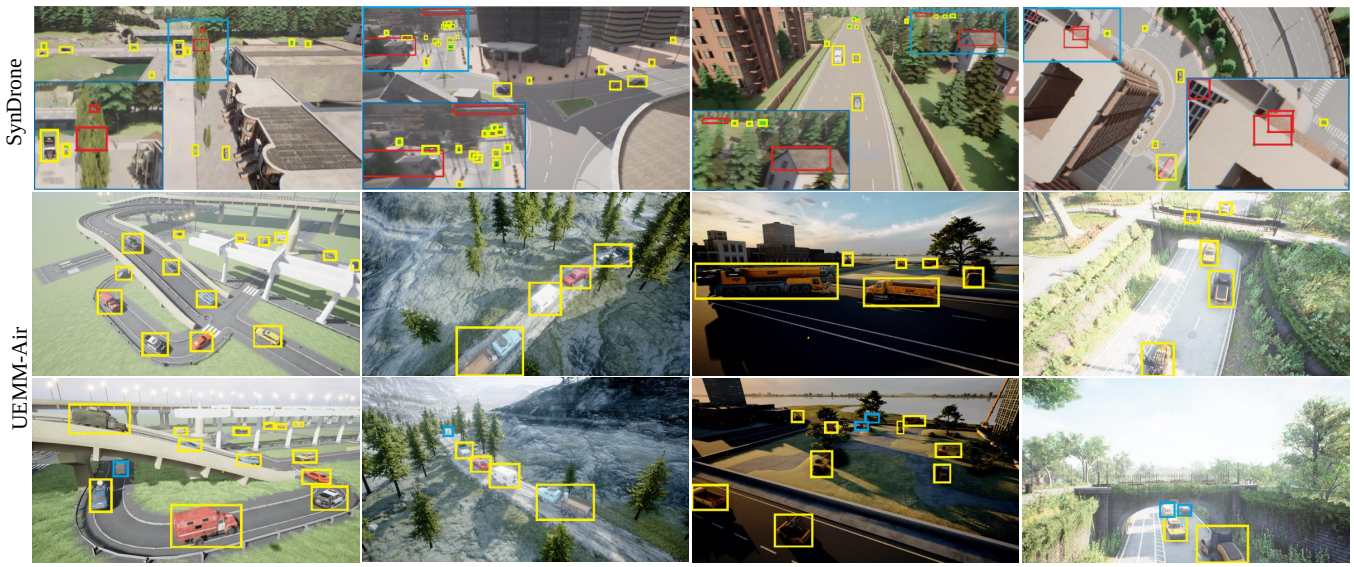


Fig. 5. Comparison of SynDrone and our UEMM-Air. Red and yellow bounding boxes indicate incorrect and correct labels, respectively. We provide two viewpoints from one scene in UEMM-Air, where blue boxes indicate originally blocked objects in the other viewpoint. SynDrone has incorrect labels where objects are visibly blocked, while UEMM-Air consistently demonstrates superior labeling accuracy, especially in challenging scenarios where objects are partially obscured.

essential for generating precise 3D reconstruction surface models.

IMU Parameters: IMU parameters encompass dynamic state data, GPS information, flying altitude and timestamps. The dynamic state parameters consist of attitude angle, linear velocity, body angular velocity, linear acceleration, and collective angular acceleration.

IMU parameters comprise real-time attitude details and UAV position coordinates. In multimodal tasks, the current flying altitude of the UAV can be utilized to assist in determining the scale information of the object. For example, the current frame's flight posture is beneficial for the model to predict the next frame's object location, especially in tasks like video object detection or object tracking. Additionally, the UAV's GPS information can also be employed for post-detection localization tasks.

C. Automatic Image Segmentation and Annotation

Most of the existing UAV-based datasets are manually annotated. Manual image annotation faces challenges in terms of accuracy and efficiency, especially when dealing with a large number of labels or low-resolution images.

To avoid manual annotation, the SynDrone [65] dataset employs an automatic image labeling algorithm. They derive the absolute coordinates of UAV and vehicles from Unreal Engine, then obtain the bounding boxes of the objects by analyzing their relative position. However, this strategy causes some incorrect annotations where objects are visibly blocked but their coordinates are still marked on the image, as illustrated in Fig. 5.

In order to alleviate the problem of mislabeling in the SynDrone dataset, we propose a heuristic automatic image annotation algorithm. It makes full use of semantic and distance information from segmentation and depth images to avoid labeling visually blocked objects and mislabeling overlapped ones. Our approach is illustrated in Fig. 3.

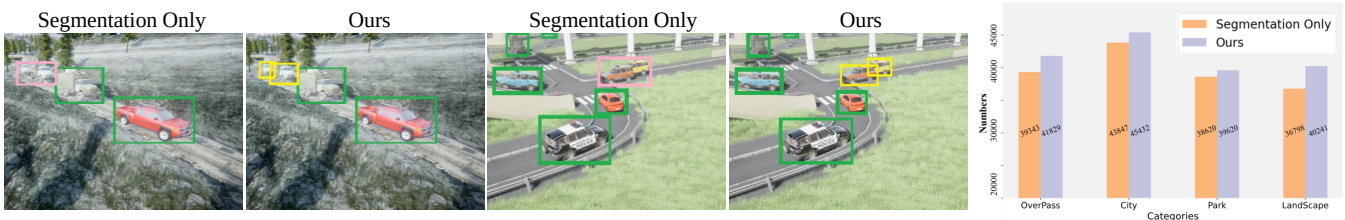
Employing the AirSim simulator, we assign the same color label to the same class of objects in the Unreal Engine environment. For each class, we convert contour detection on objects into bounding boxes and get the initial annotation. However, this step cannot recognize objects of the same category that are overlapped in the segmentation image and will mark them as one object.

To avoid mislabeling visually overlapped objects, we utilize depth images where pixel value represents the distance from the object to the camera plane to perform a secondary annotation. Intuitively, depth values mildly change on each object and a depth value jump indicates multiple objects existences. Therefore, overlapped objects can be correctly identified through depth observation. We detect depth mutations within segmented bounding boxes to confirm object edges, adjusting labels accordingly.

Fig. 6 presents sample annotation results for comparing our proposed algorithm with only utilizing segmentation information. It can be observed from Fig. 6 (a) that leveraging segmentation information alone can't effectively handle cases of visual overlapped (as shown by the pink box). Our approach can alleviate this issue by correctly distinguishing the two vehicles within the pink box. Fig. 6 (b) shows the numerical statistics of the annotations generated by the two methods. It can be observed that our method successfully annotates more objects, because our approach can distinguish overlapped instances and correct the annotations accurately employing depth information.

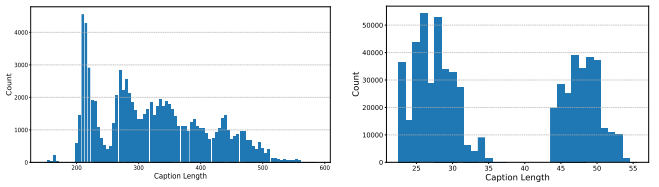
D. Cross-modality Generation

After collecting the 5 modalities mentioned in Section III-B, we also need to generate the sixth modality: text. In this



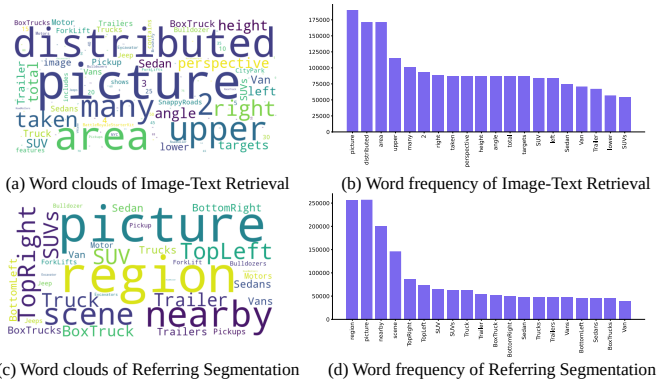
(a) Visualization comparison (b) Numerous Statistics

Fig. 6. Visual (a) and Numerical (b) Comparison with only utilizing segmentation information. The pink box indicates visually overlapped objects and the yellow box shows the corrected results. Owing to overlap rectification, our approach can generate more accurate annotations.



(a) Caption length of Image-Text Retrieval (b) Caption length of Referring Segmentation

Fig. 7. Distribution of captions length in our UEMM-Air.



(a) Word clouds of Image-Text Retrieval (b) Word frequency of Image-Text Retrieval

(c) Word clouds of Referring Segmentation (d) Word frequency of Referring Segmentation

Fig. 8. Word clouds and top 20 keywords of captions in our Dataset.

paper, we categorize text into two types: global and local, corresponding to the tasks of image-text contrastive learning and referring image segmentation, respectively.

1) Image-text Contrastive Learning: In addition to visual tasks like object detection and instance segmentation, we also hope our UEMM-Air can support vision language tasks to make UAV-based models perform zero-shot capability. In Section III-C, we already generate accurate annotations with object bounding boxes and fine-grained class names. However, such annotations cannot be directly utilized on the text encoder of CLIP. Therefore, we follow the B2C method from RemoteCLIP [66] to transfer the bboxes into a set of natural language captions. However, we notice that the captions generated by RemoteCLIP only include the quantity and position of the objects relative to the image, leading to insufficient semantic information in the captions. To alleviate this issue, we propose a new generation approach, as shown in Fig. 3. Specifically, we combine the scene and UAV parameter annotations provided by UEMM-Air to generate text descriptions that are more relevant to the UAV context. In addition, we provide more precise location information. Instead of being limited to the center and edges of the image, we utilize multiple sentences to comprehensively describe the distribution of the objects.

Ultimately, we generated 7 distinct captions for each image, resulting in a total of 840,000 descriptions. We present a visualization of the caption length distribution for our final data, as represented in Fig. 7. It can be observed that the caption length distribution shows a peak around 200, with most lengths concentrated between 200-400. Beyond 400, the frequency decreases, forming a long-tail pattern. In Fig. 8, we also provide visualizations of word clouds and the top 20 keywords of our UEMM-Air. The words exclude stop words like ”the”, ”is”, and others.

2) Referring Image Segmentation: As a prominent visual-language task, referring image segmentation enables the delineation of specific objects in the visual field through natural language descriptions. This capability is equally essential in the domain of UAVs to facilitate cross-modal interactive tasks, thereby advancing embodied intelligence in UAVs. Therefore, similar to image-text contrastive learning, we propose a method for generating textual descriptions for the referring image segmentation task. By analyzing the spatial relationships of the objects within the obtained segmentation labels, we can automate the generation of the referring captions.

Unlike referring image segmentation tasks in general domains, the UAV field often requires the segmentation of multiple objects rather than just one. For instance, it may involve segmenting a series of cars along a street. Consequently, we primarily generate descriptions in the following 3 categories:

- All objects of a specific class. *e.g.*, *All of the BoxTrucks in the image.*
- The spatial relationships of a particular object or certain objects relative to the image. *e.g.*, *Van in the bottom-left region of the image.*
- The spatial relationships of a specific object relative to other objects. *e.g.*, *The Van above the SUV in the image.*

Finally, we generated 600,000 image-mask-text pairs. Additionally, we present the statistics on captions length and word clouds in Fig. 7 and Fig. 8, respectively. Unlike the data generated in image-text contrastive learning, the length of the captions for referential tasks is shorter, as it does not require a description of the global features of the image.

3) Visualizations of Cross-modality Generation: To validate the accuracy of the generated captions, we randomly selected samples for visualization. As shown in Fig. 9, the captions for the image-text retrieval task accurately describe the number and spatial relationships of the objects, as well as the UAV's

(a) Image-Text Retrieval (b) Referring Image Segmentation

Fig. 9. The visualizations for the two types of cross-modal generation. (a) In the image-text retrieval task, the text describes the global information of the image. (b) In the referring image segmentation task, the text focuses on describing the details of local objects.

TABLE II
THE RESULTS OF THE 5-FOLD CROSS-VALIDATION EXPERIMENT. FOLD 1-5 REPRESENT THE FIVE RANDOMLY PARTITIONED SUB-DATASETS.

Fold	Train	Valid	AP_{50}	AP_{75}
1	1,2,3,4	5	63.9%	52.6%
2	1,3,4,5	4	62.6%	53.4%
3	1,2,4,5	3	63.1%	52.8%
4	1,2,3,5	2	62.3%	53.7%
5	2,3,4,5	1	64.3%	53.3%

TABLE III
UEMM-AIR TRANSFERABILITY VALIDATIONS. WE SELECTED SYNDRONE DATASET FOR COMPARISON.

Method	Fine-tuned	Pre-trained	AP_{50}	AP_{75}
YOLOv11 [67]	VisDrone	SynDrone	25.6%	16.0%
		UEMM-Air	**28.3%**	**17.7%**
	UAV-DT	SynDrone	85.4%	55.6%
		UEMM-Air	**86.2%**	**56.1%**
FasterRCNN [44]	VisDrone	SynDrone	5.1%	1.1%
		UEMM-Air	**5.5%**	**2.3%**
	UAV-DT	SynDrone	48.0%	9.5%
		UEMM-Air	**53.8%**	**14.7%**

height and angle information. The captions for the referring image segmentation specifically describe the location of a particular object and successfully generate the mask for the object.

IV. BENCHMARK AND EXPERIMENTS

A. Experimental Setup

1) Benchmark Models: We adopted YOLOv11 [67], RT-DETR [68] and D-FINE [69] as the general object detection baseline models. In the multi-modal object detection experiments, we designed a dual-path multi-modal detector with mid-level feature fusion. We utilized YOLOv7-L [70] as the base detector, with two separate backbone networks to extract features from two modalities. We also designed a feature fusion module that utilizes Coordinate Attention (CA) [71]. Specifically, we first directly concatenated the features of two modalities and employed Coordinate Attention to fuse them simultaneously in terms of channel and spatial information. The fused features were then entered into the neck part of the detector to complete the remaining detection tasks.

In image-text contrastive learning experiments, we employed OpenAI CLIP [27] and RemoteCLIP [66] as benchmark models. We selected four types of visual backbone architecture for the two CLIP models, ranging from ResNet-50, ViT-Base-16, ViT-Base-32, and ViT-Large-14. We utilized the transformer architecture, consisting of 12 layers and 8 attention heads for text encoder. The maximum token sequence length is set to 77. The InfoNCE [72] loss operates on the [CLS] token produced by the image and text backbone.

In referring image segmentation experiments, we adopted three SoTA transformer architecture models: LAVT [73], RMSIN [74], RefSegformer [75]. Among them, RMSIN is specially designed for remote sensing scenarios.

2) Training Settings: All detection and segmentation experiments were conducted in Pytorch with a NVIDIA RTX 3090 GPU. During the model transferability verification, we set the batch size to 16 and trained for 200 epochs. In other object detection experiments, we froze the backbone network of the detector and trained for 50 epochs with a batch size of 32. All detectors were trained using an Adam optimizer [76] with a momentum of 0.937. The learning rate was initialized as 0.001 with a cosine decay [77]. We fix random seed to 18 to ensure the experiment's reproducibility.

The image-text contrastive learning experiments were trained on an NVIDIA RTX 3090 GPU. The training process was accelerated by employing the Adam optimizer [76]. The learning rate was set to 7e-5, 4e-5, and 1e-4, respectively, for ResNet-50, ViT-Base-32, and ViT-Large-14 models, and the corresponding batch size was set to 256, 128, and 28, respectively.

The referring image Segmentation experiments were deployed in 4 $\times$ NVIDIA RTX 4090 GPUs. The initial learning rate was set to 0.0003, with a batch size of 6 per GPU and the Adam [76] optimizer. The training was conducted for a total of 40 epochs. The input size of images was set to 480 $\times$ 480.

B. Evaluation on Automatic Annotation Algorithm

Considering that our labeling algorithm is auto-generated, it is necessary to validate the reliability of the labels we generate. In this section, we demonstrated the effectiveness of our labels

Fig. 10. Comparison with Manual Annotations. Different colored boxes represent different categories. The generated annotations are basically consistent with manual annotations. Furthermore, the generated labels exhibit better fidelity compared to manual labels.

TABLE IV
COMPARISON OF FINE-GRAINED OBJECT DETECTION RESULTS WITH VARIOUS DETECTORS.

Method	Version	Input Size	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	#Params (M).	BFLOPs	FPS (bs=1)
YOLOv11 [67]	n	640×640	50.8%	68.9%	54.9%	22.0%	80.6%	94.5%	**2.6**	**6.3**	**263.2**
	s	640×640	56.3%	73.8%	61.5%	29.2%	84.4%	96.3%	9.4	21.3	212.8
	m	640×640	57.0%	74.8%	62.3%	30.4%	84.6%	96.0%	20.0	67.7	185.2
	l	640×640	58.6%	76.0%	64.6%	32.8%	85.5%	96.8%	25.3	86.6	158.7
	x	640×640	60.0%	76.8%	66.1%	34.7%	86.5%	97.3%	56.8	194.5	94.3
RT-DETR [68]	l	640×640	38.4%	60.3%	40.3%	11.0%	64.1%	84.7%	32.0	103.5	90.9
	x	640×640	40.1%	73.7%	38.4%	15.7%	63.1%	82.4%	65.5	222.5	56.2
D-FINE [69]	n	640×640	63.5%	82.5%	68.7%	41.9%	86.3%	95.2%	7.1	3.7	44.1
	s	640×640	71.8%	90.3%	78.0%	54.0%	89.8%	96.2%	24.9	10.2	38.8
	m	640×640	73.2%	90.7%	78.8%	55.6%	91.4%	97.1%	56.4	19.2	31.4
	l	640×640	74.3%	91.3%	80.0%	57.2%	91.8%	97.3%	90.7	30.7	23.1
	x	640×640	**75.8%**	**92.4%**	**81.8%**	**59.5%**	**92.5%**	**97.5%**	61.55	202.2	18.8

TABLE V
COMPARISON OF SEGMENTATION RESULTS WITH VARIOUS METHODS.

Method	Version	Input_Size	mAP	AP_{50}	AP_{75}
YOLOv11	n	640×640	31.4%	58.8%	30.2%
	s	640×640	34.8%	64.3%	33.9%
	m	640×640	37.7%	67.9%	37.1%
	l	640×640	38.3%	68.6%	37.9%
	x	640×640	**39.1%**	**69.5%**	**38.9%**

TABLE VI
COMPARISON OF THE TRAINING OVER DIFFERENT MODALITIES.

Modality	mAP	AP_{50}	AP_{75}
RGB	50.3%	68.4%	53.2%
RGB + Seg	55.6%	75.3%	54.1%
RGB + Surface	55.8%	74.3%	54.4%
RGB + Depth	53.5%	73.0%	54.2%
RGB + Seg + Surface + Depth	**57.3%**	**78.2%**	**58.0%**

through cross-validation experiments and visual comparisons with manual labels.

1) Verification of the Annotations' Reliability: We randomly divided the dataset into 5 parts and conducted five-fold cross-validation experiments on the YOLOv11 model. The experimental results are demonstrated in Table II. By sequentially using different sets of 4 parts as the training set and the remaining part as the validation set, we observed that the results of the 5-fold cross-validation were quite similar. The lowest AP_{50} was 62.3%, and the highest was 64.3%, with a range of 2.0%. This result indicates that the annotations we generated are consistent in their distribution.

2) Comparison with Manual Annotations: We randomly selected a subset of images for manual annotation and then visually compared them with our automatic annotations. As presented in Fig. 10, we visualized our generated labels and manually annotated labels separately for comparison. It can be observed that our annotations are almost identical in position to the manual labels. Moreover, our annotation algorithm has some advantages in labeling small objects. We found that when objects are far away, manual annotations may contain errors due to the smaller scale of the objects. For example, manual annotations may not be as closely aligned with the edges of the objects as our generated annotations. It will introduce more foreground information, which could impact the model's accuracy.

C. Evaluation on Object Detection

In this section, we selected several mainstream detectors and conducted experiments on object detection tasks at coarse-grained and fine-grained labels and multi-modal object detection tasks. Then we conducted analysis based on the model

TABLE VII
CROSS-MODAL RETRIEVAL PERFORMANCE ON UEMM-AIR.

Method	publication	Image Backbone	Text Backbone	Image To Text				Text To Image				Mean Recall
				R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	
CLIP [27]	ICML2021	ResNet50	Transformer	6.79	27.38	44.05	26.07	4.57	19.33	32.41	18.77	22.42
		ViT-B-16		8.67	30.77	47.35	28.93	5.96	22.48	35.71	21.38	25.15
		ViT-B-32		12.80	43.51	**62.72**	**39.59**	8.59	**31.09**	**47.82**	**29.17**	**34.38**
		ViT-L-14		**12.81**	**44.62**	61.99	39.32	**8.68**	30.68	46.69	29.08	34.16
RemoteCLIP [66]	TGRS2024	ResNet50	Transformer	7.02	27.34	43.72	26.03	4.56	19.17	32.33	18.69	22.36
		ViT-B-32		12.24	41.58	60.77	38.19	8.39	30.84	47.52	28.92	33.56
		ViT-L-14		12.31	41.21	60.84	38.12	8.36	30.91	46.63	28.46	33.29

TABLE VIII
REFERRING IMAGE SEGMENTATION PERFORMANCE ON UEMM-AIR.

Methods	Publication	Image Backbone	Text Backbone	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	oIoU	mIoU
LAVT [73]	CVPR22	Swin-B	BERT	53.17%	42.42%	35.47%	23.69%	8.03%	63.72%	51.09%
RMSIN [74]	CVPR24	Swin-B	BERT	**57.40%**	**49.20%**	**38.80%**	25.40%	8.60%	**65.82%**	**51.97%**
RefSegformer [75]	IEEE TIP24	Swin-B	BERT	55.28%	48.55%	38.01%	**26.04%**	**8.62%**	64.99%	51.23%

performance and experimental results. These experimental results will serve as the baseline results of our dataset for future research.

1) Transferability Verification: To demonstrate the advantage of our UEMM-Air on model transferability, We pre-trained two detectors utilizing SynDrone and UEMM-Air, respectively. Then we subsequently fine-tuned the models on the VisDrone and UAVDT datasets. The experimental results are presented in Table III. While the number of images is smaller than SynDrone ($20k$ & $60k$), the model pre-trained on the UEMM-Air dataset demonstrates stronger generalization performance on real-world scenario data. For example, after obtaining pre-trained weights on UEMM-Air and SynDrone datasets, we fine-tuned the YOLOv8 model on the VisDrone dataset. The model pre-trained on UEMM-Air demonstrated a 2.7% improvement in AP_{50} and a 1.7% improvement in AP_{75}. This might be attributed to the provision of more accurate annotations, more categories, and more diverse scenarios in UEMM-Air for the model pre-training process.

2) General Object Detection: We trained several state-of-the-art detection frameworks, including YOLOv11 [67], RT-DETR [47], and D-FINE [69], utilizing our UEMM-Air. Experimental outcomes are presented in Table IV. In terms of detection accuracy, the D-FINE-x model achieved the best performance with a mean Average Precision (mAP) of 75.8%. Similar to performances on other UAV-OD datasets, the detection accuracy for small objects was 59.5%, which is significantly lower compared to 97.5% for large objects. This indicates that our dataset poses significant challenges for small object detection, providing valuable insights for researchers aiming to tackle the difficulties associated with small object detection. Additionally, considering the real-time requirements of UAVs, we also tested the inference time metrics. The YOLOv11-n, as a lightweight model, achieved 263.2 FPS on an RTX 3090 GPU, with only 2.6M parameters.

3) Multi-modal Object Detection: In Table VI, we conducted mid-level fusion experiments for multi-modal object detection with RGB modality and the other three modalities. The model fusion of RGB with segmentation modality achieved the best performance on AP_{50}, surpassing the baseline model (RGB only) by 6.9%. The fusion of RGB with surface normal modality achieved the best performance on AP_{75}, surpassing the baseline model by 1.2%. However, fusion with depth modalities resulted in the lowest performance. This could be due to the distinct features of object positions in segmentation modality and the detailed texture features in surface normal, both containing more effective information compared to depth. We also combined 4 modalities in our experiments, achieving a mAP of 57.3%, which is an improvement of 7% compared to utilizing the RGB modality alone.

D. Evaluation on Instance Segmentation

We selected the YOLOv11 framework to conduct instance segmentation experiments, as shown in Table V. It provides benchmark results for our dataset. As the largest scale model, YOLOv11-x achieved the best accuracy, with a mAP of 39.1%, AP_{50} of 69.5%, and AP_{75} of 38.9%. Compared to object detection, instance segmentation is a more challenging dense prediction task, resulting in a relatively lower mAP.

E. Evaluation on Image-Text Contrastive Learning

Table VII presents the performance of CLIP and RemoteCLIP in image-text retrieval on our dataset. We report the retrieval recall of top-1 (R@1), top-5 (R@5), top-10 (R@10), and the mean recall of these values. From Table VII, it can be observed that the original OpenAI CLIP performs better, achieving the highest Mean Recall of 34.38. The versions using ViT-B-32 and ViT-L-14 as visual backbones perform similarly across several metrics, but ViT-B-32 demonstrates superior average performance in both retrieval tasks. Additionally, it is noteworthy that RemoteCLIP, as the CLIP model for remote sensing, performs worse than the original CLIP. This may be due to the reduced generalization capability of RemoteCLIP after fine-tuning on remote sensing satellite data.

F. Evaluation on Referring Segmentation

We conducted experiments with three state-of-the-art referring image segmentation models on our UEMM-Air, with the results summarized in Table VIII. The results indicate that RMSIN exhibits the best performance on both the mIoU and oIoU average performance metrics. It can be attributed to RMSIN's architecture is specifically designed for aerial perspectives, making it more conducive to feature learning from a UAV's viewpoint. Additionally, RMSIN's metrics at Pr@0.8 and Pr@0.9 are slightly lower than those of RefSegformer, aligning with the performance differences reported in the original paper for the two models. The phenomenon indirectly validates the high accuracy of the annotations generated automatically in our dataset.

V. Conclusion

In this paper, we release a synthetic UAV-based environmental perception dataset, named UEMM-Air. Our work achieves three main breakthroughs: Firstly, to the best of our knowledge, UEMM-Air is the largest in terms of data scale, featuring the most paired modalities and the highest number of task types. Secondly, we design a new automatic annotation method, enhancing the accuracy of annotations by employing segmentation and depth images. Then, we generate a large number of text descriptions utilizing the annotations, further enriching our dataset with text modality. Finally, we provide benchmark results across multiple tasks, thereby expanding the breadth of tasks in the field of UAV-based environmental perception. We will continue to build new simulated scenarios in the future to expand the scale and number of modalities in our dataset, supporting research on UAV-based multi-modal perception tasks.

Acknowledge

This work was supported in part by the National Natural Science Foundation of China under Grant 62372155 and Grant 62302149, in part by the Postgraduate Research and Practice Innovation Program of Jiangsu Province under Grant SJCX24_0183, in part by the Fundamental Research Funds for the Central Universities under Grant B240201077, in part by the Aeronautical Science Fund under Grant 2022Z071108001, in part by the Qinglan Project of Jiangsu Province, and in part by Changzhou Science and Technology Bureau Project No. 20231313.

References

[1] H. Huang, Y. Tang, Z. Tan, J. Zhuang, C. Hou, W. Chen, and J. Ren, "Object-based attention mechanism for color calibration of uav remote sensing images in precision agriculture," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[2] S. Luo, H. Li, Y. Li, C. Shao, H. Shen, and L. Zhang, "An evolutionary shadow correction network and a benchmark uav dataset for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[3] E. Khankeshizadeh, A. Mohammadzadeh, H. Arefi, A. Mohsenifar, S. Pirasteh, E. Fan, H. Li, and J. Li, "A novel weighted ensemble transferred u-net based model (wetum) for post-earthquake building damage assessment from uav data: A comparison of deep learning-and machine learning-based approaches," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[4] C. Qian, H. Wu, Q. Zhang, L. Yang, and Q. Jiang, "Design and implementation of UAV formation cooperative system," in *ICAUS*, 2022.

[5] A. Srivastava and J. Prakash, "Techniques, answers, and real-world uav implementations for precision farming," *WPC*, 2023.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] K. Sharifani and M. Amini, "Machine learning and deep learning: A review of methods and applications," *World Information Technology and Engineering Journal*, vol. 10, no. 07, pp. 3897–3904, 2023.

[8] X. Yan, T. Fu, H. Lin, F. Xuan, Y. Huang, Y. Cao, H. Hu, and P. Liu, "Uav detection and tracking in urban environments using passive sensors: A survey," *Applied Sciences*, vol. 13, no. 20, p. 11320, 2023.

[9] J. Su, X. Zhu, S. Li, and W.-H. Chen, "Ai meets UAVs: A survey on AI empowered UAV perception systems for precision agriculture," *Neurocomputing*, 2023.

[10] H. Liu, Y. Yu, S. Liu, and W. Wang, "A military object detection model of uav reconnaissance image and feature visualization," *Applied Sciences*, vol. 12, no. 23, p. 12236, 2022.

[11] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[13] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[14] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang *et al.*, "Visdrone-det2019: The vision meets drone object detection in image challenge results," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[15] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 370–386.

[16] P. Mittal, R. Singh, and A. Sharma, "Deep learning-based object detection in low-altitude UAV datasets: A survey," *IVC*, 2020.

[17] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey," *IEEE GRSM*, 2021.

[18] R. A. Zitar, M. Al-Betar, M. Ryalat, and S. Kassaymehd, "A review of UAV visual detection and tracking methods," *arXiv preprint arXiv:2306.05089*, 2023.

[19] Y. Zhang, Z. Gong, W. Liu, H. Wen, P. Wan, J. Qi, X. Hu, and P. Zhong, "Empowering physical attacks with jacobian matrix regularization against vit-based detectors in uav remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[20] S. Khose, A. Pal, A. Agarwal, J. Hoffman, P. Chattopadhyay *et al.*, "Skyscenes: A synthetic dataset for aerial scene understanding," *arXiv preprint arXiv:2312.06719*, 2023.

[21] L. P. Osco, J. M. Junior, A. P. M. Ramos, L. A. de Castro Jorge, S. N. Fatholahi, J. de Andrade Silva, E. T. Matsubara, H. Pistori, W. N. Gonçalves, and J. Li, "A review on deep learning in uav remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102456, 2021.

[22] S. A. Ahmed, H. Desa, H. K. Easa, A.-S. T. Hussain, T. A. Taha, S. Q. Salih, R. A. Hasan, O. K. Ahmed, and P. S. J. Ng, "Advancements in uav image semantic segmentation: A comprehensive literature review," *Multidisciplinary Reviews*, vol. 7, no. 6, pp. 2 024 118–2 024 118, 2024.

[23] P. Blikstein, "Multimodal learning analytics," in *Proceedings of the third international conference on learning analytics and knowledge*, 2013, pp. 102–106.

[24] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[25] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6700–6713, 2022.

[26] I. Bozcan and E. Kayacan, "Au-air: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance," in *ICRA*, 2020.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable

visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
[28] J. Wu, X. Li, S. Xu, H. Yuan, H. Ding, Y. Yang, X. Li, J. Zhang, Y. Tong, X. Jiang *et al.*, "Towards open vocabulary learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
[29] Y. Qiao, C. Deng, and Q. Wu, "Referring expression comprehension: A survey of methods and datasets," *IEEE Transactions on Multimedia*, vol. 23, pp. 4426–4440, 2020.
[30] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *arXiv preprint arXiv:2306.13549*, 2023.
[31] Epic Games, "Unreal engine," https://www.unrealengine.com/.
[32] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics*, 2017. [Online]. Available: https://arxiv.org/abs/1705.05065
[33] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *European Conference on Computer Vision*, 2016.
[34] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.
[35] J. Suo, T. Wang, X. Zhang, H. Chen, W. Zhou, and W. Shi, "Hit-uav: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection," *Scientific Data*, vol. 10, no. 1, p. 227, 2023.
[36] H. Feng, L. Zhang, S. Zhang, D. Wang, X. Yang, and Z. Liu, "Rtdod: A large-scale rgb-thermal domain-incremental object detection dataset for uavs," *Image and Vision Computing*, vol. 140, p. 104856, 2023.
[37] F. Liu, L. Yao, C. Zhang, T. Wu, X. Zhang, X. Jiang, and J. Zhou, "Boost uav-based ojbect detection via scale-invariant feature disentanglement and adversarial learning," *arXiv preprint arXiv:2405.15465*, 2024.
[38] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," *ISPRS journal of photogrammetry and remote sensing*, vol. 165, pp. 108–119, 2020.
[39] X. Zhang, J. Jin, Z. Lan, C. Li, M. Fan, Y. Wang, X. Yu, and Y. Zhang, "Icenet: A semantic segmentation deep network for river ice by fusing positional and channel-wise attentive features," *Remote Sensing*, vol. 12, no. 2, p. 221, 2020.
[40] M. Rahnemoonfar, T. Chowdhury, and R. Murphy, "Rescuenet: a high resolution uav semantic segmentation dataset for natural disaster damage assessment," *Scientific data*, vol. 10, no. 1, p. 913, 2023.
[41] Z. Wang and N. Mahmoudian, "Aerial fluvial image dataset for deep semantic segmentation neural networks and its benchmarks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 4755–4766, 2023.
[42] G. Rizzoli, F. Barbato, M. Caligiuri, and P. Zanuttigh, "Syndrone – multi-modal uav dataset for urban scenarios," in *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023.
[43] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.
[44] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NeurIPS*, 2015.
[45] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016.
[46] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
[47] W. Lv, S. Xu, Y. Zhao, G. Wang, J. Wei, C. Cui, Y. Du, Q. Dang, and Y. Liu, "DETRs beat YOLOs on real-time object detection," *arXiv preprint arXiv:2304.08069*, 2023.
[48] C. Duan, Z. Wei, C. Zhang, S. Qu, and H. Wang, "Coarse-grained density map guided object detection in aerial images," in *ICCV*, 2021.
[49] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *CVPRW*, 2020.
[50] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *ICCV*, 2019.
[51] A. Meethal, E. Granger, and M. Pedersoli, "Cascaded zoom-in detector for high resolution aerial images," in *CVPR*, 2023.
[52] O. C. Koyun, R. K. Keser, I. B. Akkaya, and B. U. Töreyin, "Focus-and-detect: A small object detection framework for aerial images," *SPIC*, 2022.
[53] B. Du, Y. Huang, J. Chen, and D. Huang, "Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 435–13 444.
[54] L. Yao, F. Liu, C. Zhang, Z. Ou, and T. Wu, "Domain-invariant progressive knowledge distillation for uav-based object detection," *IEEE Geoscience and Remote Sensing Letters*, 2024.
[55] G. Zou, L. Yao, F. Liu, C. Zhang, X. Li, N. Chen, S. Xu, and J. Zhou, "Remotetrimmer: Adaptive structural pruning for remote sensing image classification," *arXiv preprint arXiv:2412.12603*, 2024.
[56] J. Xie, L. Fang, B. Zhang, J. Chanussot, and S. Li, "Super resolution guided deep network for land cover classification from remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.
[57] M. Liu, B. Fu, D. Fan, P. Zuo, S. Xie, H. He, L. Liu, L. Huang, E. Gao, and M. Zhao, "Study on transfer learning ability for classifying marsh vegetation with multi-sensor images using deeplabv3+ and hrnet deep learning algorithms," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, p. 102531, 2021.
[58] X.-W. Ye, S.-Y. Ma, Z.-X. Liu, Y. Ding, Z.-X. Li, and T. Jin, "Post-earthquake damage recognition and condition assessment of bridges using uav integrated with deep learning approach," *Structural Control and Health Monitoring*, vol. 29, no. 12, p. e3128, 2022.
[59] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.
[60] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic, "In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 607–12 616.
[61] C. Broni-Bediako, J. Xia, and N. Yokoya, "Real-time semantic segmentation: A brief survey and comparative study in remote sensing," *IEEE Geoscience and Remote Sensing Magazine*, 2023.
[62] J. Cheng, C. Deng, Y. Su, Z. An, and Q. Wang, "Methods and datasets on semantic segmentation for unmanned aerial vehicle remote sensing images: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 211, pp. 1–34, 2024.
[63] K. Pang, L. Weng, Y. Zhang, J. Liu, H. Lin, and M. Xia, "Sgbnet: An ultra light-weight network for real-time semantic segmentation of land cover," *International Journal of Remote Sensing*, vol. 43, no. 15-16, pp. 5917–5939, 2022.
[64] W. Studios, "Citybld: Procedural city creation toolkit for unreal engine 5," https://www.worldbld.com, 2024.
[65] G. Rizzoli, F. Barbato, M. Caligiuri, and P. Zanuttigh, "Syndrone-multi-modal UAV dataset for urban scenarios," in *ICCV*, 2023.
[66] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, "Remoteclip: A vision language foundation model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
[67] Ultralytics, "Ultralytics yolov11," in *https://docs.ultralytics.com*, 2024.
[68] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," *arXiv preprint arXiv:2304.08069*, 2023.
[69] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu, "D-fine: Redefine regression task in detrs as fine-grained distribution refinement," *arXiv preprint arXiv:2410.13842*, 2024.
[70] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *CVPR*, 2023.
[71] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.
[72] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
[73] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr, "Lavt: Language-aware vision transformer for referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 155–18 165.
[74] S. Liu, Y. Ma, X. Zhang, H. Wang, J. Ji, X. Sun, and R. Ji, "Rotated multi-scale interaction network for referring remote sensing image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 658–26 668.
[75] J. Wu, X. Li, X. Li, H. Ding, Y. Tong, and D. Tao, "Towards robust referring image segmentation," *IEEE-TIP*, 2024.
[76] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[77] I. Loshchilov and F. Hutter, "Stochastic gradient descent with warm restarts," in *ICLR*, 2017.